arXiv:0905.4131v1 [stat.CO] 26 May 2009

MAXIMUM LIKELIHOOD ESTIMATION FOR MARKOV CHAINS

IULIANA TEODORESCU

Abstract. A new approach for optimal estimation of Markov chains with sparse transition matrices is presented.

Contents

1. Mathematical Framework 1
2. Estimation of the Transition Probability Matrix 3
3. The Bootstrap Method 4
4. The Bootstrap Method for Finite State Markov Chains 5
5. Smoothed Estimators 7
6. Asymptotic Properties of Smoothed Estimators 8
7. Performance of Smoothed Estimators 9
8. Simulation Study Structure 11
9. Simulation Results and Conclusion 12
References 13

1. Mathematical Framework

We begin with a formal mathematical definition of a Markov chain:

Definition 1.1. Let n and d be elements of $\mathbf{N}$, such that $n \geq 1$ and $d \geq 1$. Define $\Omega = \{1, \ldots, d\}$. Consider a sequence of random variables $\{X_1, X_2, \ldots, X_n\}$ such that

$$P_{ij} = P(X_{k+1} = j | X_k = i) \tag{1.1}$$

is independent of k for all i and j in Ω. Then the sequence $\{X_1, X_2, \ldots, X_n\}$ is a Markov chain with state space Ω and transition probabilities P_{ij} for i and j in Ω.

It follows from this definition that a Markov chain with known probability distribution of the initial state is completely characterized by a $d \times d$ matrix containing the transition probabilities P_{ij},

$$P = \begin{bmatrix} P_{11} & P_{12} & \ldots & P_{1d} \\ P_{21} & P_{22} & \ldots & P_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ P_{d1} & P_{d2} & \ldots & P_{dd} \end{bmatrix}.$$

This matrix is called the *transition probability matrix*. Since the elements of row i of this matrix represent the conditional probabilities for all possible state changes from state i, they must satisfy

$$\sum_{j=1}^{d} P_{ij} = 1, \tag{1.2}$$

for all $i \in \Omega$. For a Markov chain with known transition probability matrix, the most likely state as $n \to \infty$ can be calculated as follows. Define a vector V_k so that the i^{th} element of V_k is the unconditional probability that the Markov chain is in state i at time k. Hence, $(V_k)_i = P(X_k = i)$, where $V_k' = [(V_k)_1, \ldots, (V_k)_d]$.

The probability $(V_{k+1})_i = P(X_{k+1} = i)$ can be related to the vector V_k using the Law of Total Probability,

$$(V_{k+1})_i = P(X_{k+1} = i) = \sum_{j=1}^{d} P(X_k = j)P(X_{k+1} = i|X_k = j) = \sum_{j=1}^{d} P_{ji} \cdot (V_k)_j.$$

Hence $V_{k+1} = P'V_k$. One can then use an inductive argument to establish that $V_{k+1} = (P')^k V_1$. Here V_1 is a vector of probabilities corresponding to the distribution of the initial state of the Markov chain. Hence

$$P(X_1 = j) = (V_1)_j$$

for $j = 1, \ldots, d$.

The limiting, or steady state, probabilities, if they exist, are then given by

$$\Pi^{(i)} = \lim_{n\to\infty} [(P')^n] \cdot V_1^{(i)}. \tag{1.3}$$

Since $[\Pi^{(i)}]_j = \sum_{k=1}^{d} \lim_{n\to\infty} [(P')^n]_{jk}\delta_{ik} = \lim_{n\to\infty} [(P')^n]_{ji}$, it follows that $[\Pi^{(i)}]' = [\Pi_1^{(i)}, \ldots, \Pi_d^{(i)}]$ is the i^{th} row of $P_\pi = \lim_{n\to\infty} P^n$.

Under certain conditions [14], the limit will exist and the rows of P_π will be identical. We will denote one of these rows as Π. The elements of Π correspond to the long-range probabilities that the Markov chain is in each of the states. In some instances Π can be found analytically.

Example: Consider a Markov chain with transition probability matrix

$$P = \begin{bmatrix} \frac{1+a}{2} & \frac{1-a}{2} \\ \frac{1-a}{2} & \frac{1+a}{2} \end{bmatrix},$$

where $0 \leq a < 1$. A simple induction arguments shows that

$$P^n = \begin{bmatrix} \frac{1+a^n}{2} & \frac{1-a^n}{2} \\ \frac{1-a^n}{2} & \frac{1+a^n}{2} \end{bmatrix},$$

for all integers $n \geq 1$. Since $0 \leq a < 1$, $\lim_{n\to\infty} a^n = 0$, so the limit $P_\pi = \lim_{n\to\infty} P^n$ exists and the rows of P_π are identical:

$$P_\pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

2. Estimation of the Transition Probability Matrix

In most practical cases, the transition probability matrix is unknown and it must then be estimated based on the observations. Let $X_1, X_2, \ldots, X_n$ be n consecutive observations from a Markov chain. The maximum likelihood estimator of the matrix P, which we will denote as $\widehat{P}$, is defined as follows [3]:

1) : For each state $i \in \Omega$, let n_i be the number of times that state i is observed in $X_1, X_2, \ldots, X_{n-1}$.

2) : If $n_i = 0$ (the state is not represented in the chain, except maybe for the last position), then we formally define all probabilities of transition from the state i to any state $j \neq i$ to be $0, \hat{P}_{ij} = 0$, for every $j \neq i$. Therefore, by (2), we have $\widehat{P}_{ii} = 1$.

3) : If $n_i > 0$, let n_{ij} be the number of observed consecutive transitions from state i to state j in $X_1, X_2, \ldots, X_n$. In this case, $\widehat{P}_{ij} = \frac{n_{ij}}{n_i}$, for $j = 1, \ldots, d$.

Note that the final observed state of the chain is not counted in Step 1 because we do not observe any transitions from this state. Hence, we only observe $n - 1$ transitions. Note also that the estimate $\widehat{P}$ is a valid transition probability matrix.

Since the transition probability matrix has d^2 elements, it is natural to rewrite P as a column vector with d^2 elements [1]:

$$P_v = vec(P) = \begin{bmatrix} P_{11} \\ P_{12} \\ \vdots \\ P_{1d} \\ \vdots \\ P_{d1} \\ \vdots \\ P_{dd} \end{bmatrix}$$

which allows us to concentrate on properties of the random vector P_v. This vector has d^2 elements, labeled by a two-digit index. For instance, P_{ij} is the element found on the row $k = j + (i-1)d$ of the vector $vec(P) : P_{ij} = (P_v)_k$.

The properties of the maximum likelihood estimator $\widehat{P}$ have been studied extensively [1]. In particular, $\widehat{P}$ can be shown to be asymptotically normal and consistent. The limiting probabilities computed from $\widehat{P}$ are also consistent estimates of the true limiting probabilities. These results are presented in the two theorems below.

Let $\widehat{P}_n$ be the maximum likelihood estimator corresponding to n observations $X_1, \ldots, X_n$ from a Markov chain with transition probability matrix P. Let $(\widehat{P}_v)_n$ and P_v be the vector forms of $\widehat{P}_n$ and P, respectively. The following theorem describes the asymptotic properties of the vector $(\widehat{P}_v)_n$ as $n \to \infty$.

Theorem 2.1. *As $n \to \infty$,*

$$\sqrt{n}\left[(\widehat{P}_v)_n - P_v\right] \xrightarrow{w} N(O, \Sigma_P), \tag{2.1}$$

where Σ_P is given by

$$(\Sigma_P)_{(ij,kl)} = \delta_{ik} P_{ij}(\delta_{jl} - P_{il}). \tag{2.2}$$

Here, Σ_P is a square $d^2 \times d^2$ matrix. The matrix element displayed corresponds to the row $j + (i-1)d$ and the column $l + (k-1)d$.

Now assume that for all integers $n > 0$, the limit $\lim_{m\to\infty} [\widehat{P}_n]^m$ exists and has all rows identical. Denote by $\widehat{\Pi}_n$ and Π the steady-state probabilities corresponding to $\widehat{P}_n$ and P, respectively. The following theorem establishes the consistency of the estimates of steady-state probabilities.

Theorem 2.2. *For all i, $(\widehat{\Pi}_n)_i \to (\Pi)_i$, with probability 1, as $n \to \infty$, where $(\widehat{\Pi}_n)_i$ and $(\Pi)_i$ are the i^{th} elements of $\widehat{\Pi}_n$ and Π respectively.*

These results provide an asymptotic justification of the use of $\widehat{P}$ to estimate P. When the sample size is not sufficiently large, the asymptotic results given in previous results may not hold. In these cases, the bootstrap method, which is outlined in the next section, can be used to find approximate results corresponding to those given above.

3. The Bootstrap Method

Let X be a random variable with distribution function F and let $\mathbf{X} = (x_1, \ldots, x_n)'$ be an observed sample from F. Suppose $R(\mathbf{X}, F)$ is a statistical quantity that depends in general on both the unknown distribution F and on the sample $\mathbf{X}$. For example, $R(\mathbf{X}, F)$ could be an estimator of an unknown parameter. If F is unknown, then the exact distribution of the random variable $R(\mathbf{X}, F)$ is generally unknown.

In 1979, Efron [5] proposed the *bootstrap* method to nonparametrically estimate the distribution of $R(\mathbf{X}, F)$. The method consists of the following three steps:

- **(i):** From the observed sample $\mathbf{X}$, use the empirical distribution function, $\widehat{F}_n$, as an estimate of the probability function F. The empirical distribution function is defined by $\widehat{F}_n(x) = \frac{n(x)}{n}$, where $n(x)$ is the number of values x_i in $\mathbf{X}$ that are less than or equal to x.
- **(ii):** Draw B samples of size n from $\widehat{F}_n$ conditional on $\mathbf{X}$. Denote these as $\mathbf{X}_j^*$, for $j = 1, \ldots, B$.
- **(iii):** For each sample $\mathbf{X}_j^*$, compute $R_j^* = R(\mathbf{X}_j^*, \widehat{F}_n)$ and approximate the distribution of $R(\mathbf{X}, F)$ with the empirical distribution of $R_1^*, \ldots, R_B^*$.

The samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^*$ are called resamples and the empirical distribution of $R_1^*, \ldots, R_B^*$ is called the bootstrap estimate of the distribution of R, or simply the bootstrap distribution of R^*.

The *bootstrap principle* states that the empirical distribution of $R_1^*, \ldots, R_B^*$ is a good approximation to the true distribution of $R(\mathbf{X}, F)$. Several authors have proven that the approximation is asymptotically valid for a large number of statistics of interest, and underlying populations, under some regularity conditions. See [4] and [6].

In [11], Kulperger and Prakasa Rao studied the applicability of the bootstrap method to the problem of estimating properties of Markov chains. Working under certain assumptions, they proved the following Central Limit Theorem for the bootstrap maximum likelihood estimator matrices.

Let $X_1, \ldots, X_n$ be n observations from a Markov chain with transition probability matrix P and let $\widehat{P}_n$ be the maximum likelihood estimator of P computed

on the sample. Generate a bootstrap chain, $X_1^*, \ldots, X_n^*$, by generating a Markov chain with transition probability matrix $\widehat{P}_n$, conditional on $X_1, \ldots, X_n$. Denote the maximum likelihood estimator for the bootstrap chain by $\widehat{P}_n^*$. Let $(\widehat{P}_v^*)_n$ and $(\widehat{P}_v)_n$ be the vector forms of $\widehat{P}_n^*$ and $\widehat{P}_n$, respectively.

Theorem 3.1. *There is a sequence $N_n \in \mathbb{N}$, such that*

$$\sqrt{N_n}\left[(\widehat{P}_v^*)_n - (\widehat{P}_v)_n\right] \xrightarrow{w} N(0, \Sigma_P), \tag{3.1}$$

as $n \to \infty$ and $N_n \to \infty$, where

$$(\Sigma_P)_{(ij,kl)} = \delta_{ik} P_{ij}(\delta_{jl} - P_{il}). \tag{3.2}$$

This result indicates that the distribution of the bootstrap maximum likelihood estimator has similar asymptotic behavior as the distribution of the maximum likelihood estimator.

4. The Bootstrap Method for Finite State Markov Chains

When applied to the problem of estimating Markov chains, the bootstrap method consists of computing $\widehat{P}$ from the original chain, and then generating B additional samples based on $\widehat{P}$. A uniform probability distribution for the initial state is used. For each of these resamples, a maximum likelihood estimator $\widehat{P}_i^*$, $i = 1, \ldots, B$ is computed. Based on the vector sample $(\widehat{P}_v^*)_1, \ldots, (\widehat{P}_v^*)_B$, estimators for $E(P_v)$ and $Cov(P_v)$ can be computed as follows:

$$\widehat{E(P_v)} = \frac{1}{B}\sum_{k=1}^{B}(\widehat{P}_v^*)_k,$$

$$\widehat{Cov(P_v)} = \frac{1}{B-1}\sum_{k=1}^{B}\left[(\widehat{P}_v^*)_k - \widehat{E(P_v)}\right]\cdot\left[(\widehat{P}_v^*)_k - \widehat{E(P_v)}\right]',$$

where $\widehat{Cov(P_v)}$ is a square matrix of dimension $d^2 \times d^2$.

The empirical distribution function for each element $(P_v)_{ij}$ of the vector P_v can also be computed, based on the sample $[(\widehat{P}_v^*)_1]_{ij}, \ldots, [(\widehat{P}_v^*)_B]_{ij}$. Denote this function by $\widehat{F}_{ij}$. A $(1-\alpha)100\%$ confidence interval based on the percentile method of Efron (1979) (see also Reference [7]) for the element $(P_v)_{ij}$ is given by $[\widehat{F}_{ij}^{-1}(\alpha), \widehat{F}_{ij}^{-1}(1-\alpha)]$. Here, $x_L = [\widehat{F}_{ij}]^{-1}(\alpha)$ is the largest value of x such that the number of elements in the sample $[(\widehat{P}_v^*)_1]_{ij}, \ldots, [(\widehat{P}_v^*)_B]_{ij}$ that are less than x is smaller than αn. Likewise, $x_U = [\widehat{F}_{ij}]^{-1}(1-\alpha)$ is the smallest value of x such that the number of elements in the sample $[(\widehat{P}_v^*)_1]_{ij}, \ldots, [(\widehat{P}_v^*)_B]_{ij}$ that are smaller than x is larger than $(1-\alpha)n$. Specifically,

$$x_L = max\left\{x : (\widehat{F}_n)_{ij}(x) \le \alpha\right\}, \qquad x_U = min\left\{x : (\widehat{F}_n)_{ij}(x) \ge 1-\alpha\right\}.$$

The bootstrap procedure may not perform well in some circumstances. For example, under certain conditions, the matrix $\widehat{P}$ may not have a structure that is close to that of P. To illustrate one of these situations, we consider the following numerical example.

Example: Let the true transition probability matrix of a Markov chain be

TABLE 1. The ten samples generated using the transition matrix in (8)

Sample Number	Generated Sample
1	3, 4, 2, 4, 3, 4, 3, 4, 4, 1
2	2, 2, 1, 4, 4, 4, 1, 1, 4, 3
3	3, 2, 4, 3, 4, 2, 2, 4, 3, 4
4	2, 4, 4, 4, 2, 4, 4, 2, 4, 3
5	3, 2, 2, 4, 3, 4, 4, 4, 3, 4
6	4, 4, 3, 4, 3, 4, 4, 3, 4, 4
7	2, 2, 4, 4, 2, 4, 2, 3, 4, 4
8	2, 3, 4, 3, 3, 3, 4, 1, 4, 2
9	2, 4, 4, 1, 2, 3, 4, 4, 2, 3
10	1, 1, 4, 4, 1, 3, 4, 4, 4, 4

$$P = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.10 & 0.20 & 0.20 & 0.50 \\ 0.05 & 0.10 & 0.10 & 0.75 \\ 0.10 & 0.20 & 0.30 & 0.40 \end{bmatrix}. \tag{4.1}$$

Using the C code listed in Appendix A, we generated samples of length $n = 10$ from this transition matrix, using an initial distribution of $V_1 = (0.25, 0.25, 0.25, 0.25)'$. Ten such samples are listed in Table 1.

The first sample leads to the following maximum likelihood estimator $\widehat{P}$:

$$\widehat{P} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \\ 0.20 & 0.20 & 0.40 & 0.20 \end{bmatrix}$$

Note that the estimate $\widehat{P}$ is significantly different from the original matrix P. The main difference is that $\widehat{P}$ is *sparse* (has many null entries), while P is not. Therefore, many valid transitions will never occur in resamples based on the matrix $\widehat{P}$. Regardless of how many bootstrap resamples we use, the fact that all the bootstrap maximum likelihood estimators $\widehat{P}^*$ are sparse may cause the bootstrap method to give unreliable results.

Computing maximum likelihood estimators from the other samples generated from P leads again to sparse estimators, though they may differ from the one listed above. This is because the sample size chosen is relatively small compared to the total number of possible transitions ($n = 10$, for $d^2 = 16$). A maximum of only 60% of all transitions will be found in a given sample.

Another situation that leads to sparse estimators occurs when the matrix P has elements with small probabilities. In this case, it is the existence of *rare* transitions (corresponding to the small probabilities) that causes the problem. For instance, if we use the matrix $\widehat{P}$ as the true P matrix, we obtain the samples listed in Table 2.

TABLE 2. The ten samples generated using $\widehat{P}$

Sample Number	Generated Sample
1	3, 4, 1, 1, 1, 1, 1, 1, 1, 1
2	2, 4, 1, 1, 1, 1, 1, 1, 1, 1
3	3, 4, 3, 4, 3, 4, 1, 1, 1, 1
4	2, 4, 3, 4, 1, 1, 1, 1, 1, 1
5	3, 4, 1, 1, 1, 1, 1, 1, 1, 1
6	4, 4, 2, 4, 3, 4, 3, 4, 2, 4
7	2, 4, 4, 4, 1, 1, 1, 1, 1, 1
8	2, 4, 4, 3, 4, 2, 4, 1, 1, 1
9	2, 4, 4, 1, 1, 1, 1, 1, 1, 1
10	1, 1, 1, 1, 1, 1, 1, 1, 1, 1

The maximum likelihood estimator of the first sample is:

$$\widehat{P} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \\ 1.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}.$$

As indicated earlier, increasing the number of samples does not help, since all the estimators will be sparse. To avoid this from happening, one should use a non-sparse matrix to generate the resamples.

Next, we will describe a way of solving this problem, by *smoothing* the maximum likelihood estimators. This procedure replaces a sparse estimator by a modified version where all of the entries are positive.

5. Smoothed Estimators

As indicated in the previous section, a problem related to estimating the transition probability matrix from observed sample chains is the possibility that some states of the system are too rare to occur in a limited experiment. A similar result is obtained when the chain length, n, is small compared to the total number of possible transitions, d^2. In this case only a fraction of all the possible transitions will be present in any given sample. When this happens, a particular transition may not be observed in the sample, even though the probability of this transition occuring is greater than 0.

When a sparse estimator $\widehat{P}$ is obtained from the initial chain, the impact on the bootstrap method is significant. If we assume that $\widehat{P}_{ij} = 0$ for some i and j, then a transition from state i to state j will never be observed in any of the resamples, even though it may be possible in the actual Markov chain. A similar problem occurs in the case of using the bootstrap on independent discrete data. In [9] and [13], the authors exhibit several examples where sparse data causes the bootstrap to perform poorly.

One solution to this problem is to increase the sample size. When a larger sample size is not feasible, the following method can be used. Since the cause of the problem is the fact that $\widehat{P}$ is sparse, we can attempt to generate the bootstrap

resamples based on a slightly different matrix, whose entries are all positive. We call this matrix the *smoothed* version of $\widehat{P}$ and denote it by $\widetilde{P}$. It is given by

$$\widetilde{P}_{ij} = \frac{1}{\omega}[\widehat{P}_{ij} + n^{-u}], \tag{5.1}$$

where

$$\omega = \sum_{j=1}^{d}[\widehat{P}_{ij} + n^{-u}] = \sum_{j=1}^{d}\widehat{P}_{ij} + \sum_{j=1}^{d} n^{-u} = 1 + n^{-u}d,$$

and $u > 0$ is a positive smoothing parameter.

The form of this smoothed matrix is based on simple smoothers that are used for multinomial distributions. See, for example, [8] and [16]. Note that from the definition, we obtain

$$\sum_{j=1}^{d}\widetilde{P}_{ij} = \frac{1 + dn^{-u}}{\omega} = 1, \quad \text{for all} \quad i = 1, \ldots, d,$$

so that $\widetilde{P}$ is a valid transition probability matrix.

The choice of the smoothing parameter u presents some difficulty. It is technically possible to specify a performance criterion for $\tilde{P}$ in terms of some measure of the performance of the resulting bootstrapping method. The parameter u could then be chosen to optimize this criterion. However, it is unlikely that such a method would be feasible in practice, and is well beyond the scope of this study. Nevertheless, we will justify some general properties that u should follow. These will ensure that the smoothing does not asymptotically affect the behavior of the generated Markov chains.

The criterion we choose is to select the smoothing parameter such that $\widetilde{P}$ is a consistent estimator of P at the same rate as $\widehat{P}$.

6. Asymptotic Properties of Smoothed Estimators

In the following, we consider n observations $X_1, X_2, \ldots, X_n$ from a Markov chain and establish the asymptotic properties of the smoothed estimator of the transition probability matrix. We begin by proving some general properties.

In order to study the asymptotic properties of estimators, we must introduce the following equivalence relation for matrices.

Let $\{P_n\}$ and $\{R_n\}$ be two sequences of $d \times d$ matrices, for $n = 1, 2, \ldots$. Suppose there is an $r > 0$ such that the sequence $n^r(E_n)_{ij} = n^r(P_n - R_n)_{ij}$ has the property that it remains bounded as $n \to \infty$ for all $i, j = 1, \ldots, d$. Then as $n \to \infty$

$$P_n = R_n + O(n^{-r}). \tag{6.1}$$

Here, $O(n^{-r})$ represents any sequence of matrices properly bounded.

The following theorem describes the asymptotic consistency property of the smoothed estimator defined earlier.

Theorem 6.1. *Suppose $\widehat{P} = P + O(n^{-k})$ as $n \to \infty$ for some $k > 0$. Then $\widetilde{P} = P + O(n^{-k})$ as $n \to \infty$ as long as $u \geq k$.*

Proof:
Consider the function $f(x) = (1+x)^{-1}$. A Taylor expansion of f around $x = 0$ is

$$f(x) = 1 + O(x) \text{ as } x \to 0.$$

We can rewrite $\omega^{-1} = f(n^{-u}d)$ so that

$$\omega^{-1} = 1 + O(n^{-u}), \text{ as } n \to \infty, \tag{6.2}$$

since $n^{-u}d$ remains bounded as $n \to \infty$ for fixed integer $d \geq 1$.
Now computing

$$n^u[n^{-u}\omega^{-1}] = \omega^{-1} = 1 + O(n^{-u}),$$

which by definition [15] remains bounded as $n \to \infty$, so

$$n^{-u}\omega^{-1} = O(n^{-u}). \tag{6.3}$$

In matrix notation, this result can be rewritten as

$$\widetilde{P} = \omega^{-1}\widehat{P} + n^{-u}\omega^{-1}J, \tag{6.4}$$

where J is a $d \times d$ matrix with all entries equal to 1. We conclude that:

$$\widetilde{P} = \widehat{P} + a_n\widehat{P} + b_nJ, \tag{6.5}$$

where the sequences $n^u a_n$ and $n^u b_n$ remain bounded as $n \to \infty$. Then for all $i, j = 1, \ldots, d$, $0 \leq \widehat{P}_{ij} \leq 1$ and $J_{ij} = 1$, so $n^u[a_n\widehat{P}_{ij} + b_nJ_{ij}]$ remains bounded as $n \to \infty$. Therefore,

$$\widetilde{P} = \widehat{P} + O(n^{-u}), \qquad \text{as } n \to \infty. \tag{6.6}$$

Since $\widehat{P} = P + O(n^{-k})$, we can write

$$\widetilde{P} = P + A_n + B_n, \tag{6.7}$$

where $n^k(A_n)_{ij}$ and $n^u(B_n)_{ij}$ remain bounded as $n \to \infty$. Then for all $k \leq u$, $[n^k(A_n)_{ij} + n^{k-u}n^u(B_n)_{ij}]$ remains bounded as $n \to \infty$, so,

$$\widetilde{P} = P + O(n^{-k}), \tag{6.8}$$

as long as $k \leq u$.

As shown in [1] and [3], the exponent k is usually equal to 0.5. Therefore, any choice of u such that $u \geq 0.5$ will ensure that $\widetilde{P}_n$ preserves the asymptotic consistency property of $\widehat{P}_n$.

7. Performance of Smoothed Estimators

To compare the performance of the smoothed and unsmoothed estimators, we present two Examples.

Example: In this example we explore the behavior of the bootstrap bias estimator using $\widehat{P}$ and $\widetilde{P}$. We use the transition probability matrix from the example given in (8). The true probability matrix is:

$$P = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.10 & 0.20 & 0.20 & 0.50 \\ 0.05 & 0.10 & 0.10 & 0.75 \\ 0.10 & 0.20 & 0.30 & 0.40 \end{bmatrix}.$$

Using a chain generated from P, with uniform distribution probability for the initial state, the following maximum likelihood estimator is computed:

$$\widehat{P} = \begin{bmatrix} 0.111111 & 0.222222 & 0.222222 & 0.444444 \\ 0.142857 & 0.142857 & 0.357143 & 0.357143 \\ 0.000000 & 0.037037 & 0.185185 & 0.777778 \\ 0.122449 & 0.183673 & 0.285714 & 0.408163 \end{bmatrix}.$$

With smoothing parameter $u = 0.5$ the smoothed maximum likelihood estimator is

$$\widetilde{P} = \begin{bmatrix} 0.150794 & 0.230159 & 0.230159 & 0.388889 \\ 0.173469 & 0.173469 & 0.326531 & 0.326531 \\ 0.071429 & 0.097884 & 0.203704 & 0.626984 \\ 0.158892 & 0.202624 & 0.275510 & 0.362974 \end{bmatrix}.$$

After applying the bootstrap method, with $B = 1000$, the average estimator computed from the samples based on the unsmoothed matrix is found to be

$$\overline{\widehat{P}} = \begin{bmatrix} 0.099799 & 0.220662 & 0.233852 & 0.445688 \\ 0.145533 & 0.139222 & 0.359566 & 0.355678 \\ 0.000000 & 0.035484 & 0.177676 & 0.786840 \\ 0.121927 & 0.180267 & 0.287282 & 0.410525 \end{bmatrix}$$

The average computed from the sample based on the smoothed estimator is given by

$$\overline{\widetilde{P}} = \begin{bmatrix} 0.171886 & 0.240630 & 0.241419 & 0.346065 \\ 0.196326 & 0.191570 & 0.307840 & 0.304264 \\ 0.122615 & 0.141866 & 0.212862 & 0.522657 \\ 0.183719 & 0.214623 & 0.270342 & 0.331316 \end{bmatrix}.$$

As we can see, the smoothed estimator contains some information about the low-probability transitions of the system, while the standard maximum likelihood estimator does not. In particular, the element corresponding to the transition $3 \to 1$, which has the lowest probability for this chain, is strictly zero in the average maximum likelihood estimator, but not in the smoothed version. Since the average is computed from non-negative numbers, it follows that $(\widehat{P}^*)_{31} = 0$ for all the resamples based on $\widehat{P}$. The bootstrap method based on $\widehat{P}$ leads to the conclusion that the transition $3 \to 1$ is not allowed in this chain.

The bootstrap method based on $\widetilde{P}$ does not lead to the same conclusion, as all the elements of $\overline{\widetilde{P}}$ are positive. While $(\overline{\widetilde{P}})_{31}$ is not very close to the true value 0.05, the confidence interval for this element predicted by the bootstrap method based on $\widetilde{P}$ may have good coverage properties. The same conclusion holds for other statistical inference quantities. A simulation study of the coverage properties of the bootstrap confidence intervals is presented in the next section.

TABLE 3. Asymptotic Behavior of $\sqrt{n}(\widehat{P}_n - P)$

Sample Size	$\sqrt{n}(\widehat{P}_n - P)$			
$n = 50$	-0.353553	-0.353553	-0.353553	1.060660
	0.303046	0.606092	-1.414214	0.505076
	-0.353553	-0.707107	0.380750	0.679910
	0.176777	-0.235702	0.530330	-0.471404
$n = 100$	-1.388889	-0.277778	-0.277778	1.944444
	0.428571	-0.571429	1.571429	-1.428571
	-0.500000	-0.629630	0.851852	0.277778
	0.224490	-0.163265	-0.142857	0.081633
$n = 500$	-0.356819	0.118940	0.118940	0.118940
	1.000346	-1.529941	2.000692	-1.471097
	-0.396722	-0.252459	-0.432787	1.081969
	-0.372678	-0.656623	0.301691	0.727609
$n = 1000$	-1.956855	-0.704468	1.800307	0.861016
	0.866102	-1.893338	0.926527	0.100710
	-0.119582	0.026574	-0.637770	0.730779
	0.044008	-0.792141	0.006286	0.741846
$n = 10,000$	-3.185526	1.503569	0.891948	0.790009
	0.154021	-0.091273	0.992584	-1.055333
	-0.764507	-0.681914	0.758153	0.688267
	-0.028548	-1.158239	0.261009	0.925773

Example: In this example we explore the asymptotic behavior of $\widehat{P}_n$ and $\widetilde{P}_n$ as $n \to \infty$. The matrix given in (8) is the true transition probability matrix of the system. Single samples of size 50, 100, 500, 1000 and 10,000 were generated based on P. For each sample, the estimators $\widehat{P}_n$ and $\widetilde{P}_n$ were computed. The matrices $\sqrt{n}(\widehat{P}_n - P)$ and $\sqrt{n}(\widetilde{P}_n - P)$ were then calculated. The results are listed in Tables 3 and 4.

The matrices listed in Tables 3 and 4 indicate that for each $i, j = 1, \ldots, d$, $[\sqrt{n}(\widehat{P}_n - P)]_{ij}$ and $[\sqrt{n}(\widetilde{P}_n - P)]_{ij}$ remain bounded as $n \to \infty$ and that they are of the same order of magnitude. In fact, simulations up to $n = 1,000,000$ indicate exactly the same result. This example demonstrates by direct computation that $\widehat{P}_n - P = O(n^{-0.5})$ and $\widetilde{P}_n - P = O(n^{-0.5})$

8. Simulation Study Structure

The goal of this simulation study is to perform a quantitative comparison between the performance of the bootstrap method based on the maximum likelihood estimator and its smoothed version. The true transition probability matrix P is known. To ensure that the structure of P does not unduly influence the results, two different transition probability matrices were used:

$$P_{\mathrm{I}} = \begin{bmatrix} \frac{4}{10} & \frac{3}{10} & \frac{3}{10} \\ \frac{3}{10} & \frac{4}{10} & \frac{3}{10} \\ \frac{3}{10} & \frac{3}{10} & \frac{4}{10} \end{bmatrix} \text{ and } P_{\mathrm{II}} = \begin{bmatrix} \frac{2}{20} & \frac{9}{20} & \frac{9}{20} \\ \frac{9}{20} & \frac{2}{20} & \frac{9}{20} \\ \frac{9}{20} & \frac{9}{20} & \frac{2}{20} \end{bmatrix}.$$

TABLE 4. Asymptotic Behavior of $\sqrt{n}(\widetilde{P}_n - P)$

Sample Size	$\sqrt{n}(\widetilde{P}_n - P)$			
$n = 50$	−0.225814	−0.225814	−0.225814	0.677441
	0.576773	0.514849	−0.775516	−0.316107
	0.285145	−0.068409	0.626403	−0.843139
	0.496126	−0.022803	0.210981	−0.684304
$n = 100$	−0.992063	−0.198413	−0.198413	1.388889
	0.734694	−0.265306	1.265306	−1.734694
	0.214286	−0.021164	1.037037	−1.230159
	0.588921	0.026239	−0.244898	−0.370262
$n = 500$	−0.302675	0.100892	0.100892	0.100892
	1.357508	−1.128134	1.866757	−2.096130
	0.342084	0.294804	0.141840	−0.778728
	0.192828	−0.387335	0.086260	0.108245
$n = 1000$	−1.737124	−0.625364	1.598154	0.764335
	1.301476	−1.503197	1.000032	−0.798310
	0.604016	0.556217	−0.033529	−1.126703
	0.571694	−0.525651	−0.171962	0.125919
$n = 10,000$	−3.063005	1.445740	0.857643	0.759625
	0.725021	0.104546	1.146716	−1.976280
	0.034128	−0.078763	1.305917	−1.261279
	0.549473	−0.921383	0.058663	0.313245

For both of the true transition probability matrices, simulations were conducted for all the combinations of parameters $n = 25, 50, 100$ and $u = 0.5, 1.0, 2.0$ and ∞. Note that $u = \infty$ corresponds to the standard bootstrap.

Each simulation consists of the following steps:

(1) A single chain of size n is generated from the true transition probability matrix. Estimators $\widehat{P}$ and $\widetilde{P}$ are computed.
(2) The bootstrap method (as described before) is applied, using $\widehat{P}$ and $\widetilde{P}$, respectively. The number of bootstrap resamples generated is $B = 5000$.
(3) Bootstrap 90% confidence intervals for the elements P_{11} and P_{12} are computed, based on $\widehat{P}$ and $\widetilde{P}$, respectively using the bootstrap percentile method outlined previously.
(4) Steps 1-3 are repeated 1000 times and the observed coverage properties of the intervals from the two estimators are compared.

9. Simulation Results and Conclusion

The results of the small simulation study are presented in Table 5 and seem to indicate that:

(1) For almost all combinations of simulation parameters n and u, the coverage performance of the confidence intervals based on $\widetilde{P}$ is better than for the intervals based on $\widehat{P}$.

TABLE 5. The empirical coverage of the standard ($u = \infty$) and smoothed bootstrap percentile method confidence intervals for the entries P_{11} and P_{12} of P_{I} and P_{II}. The specified nominal coverage is 90%.

		P_{I}		P_{II}	
n	u	P_{11}	P_{12}	P_{11}	P_{12}
25	0.5	90.6	90.6	99.6	99.6
25	1	86.2	86.2	99.3	99.3
25	2	81.6	81.6	53.0	53.0
25	∞	81.5	85.4	53.0	85.6
50	0.5	93.1	93.1	97.8	97.8
50	1	86.8	86.8	79.3	79.3
50	2	85.4	85.4	79.6	79.6
50	∞	85.3	88.6	79.5	89.2
100	0.5	92.0	92.9	94.2	94.2
100	1	88.1	88.1	89.3	89.3
100	2	87.1	87.1	82.7	82.7
100	∞	87.0	89.1	82.4	90.2

(2) At fixed chain length, n, increasing the smoothing parameter u leads to narrower confidence intervals, with lower coverage performance.
(3) Increasing the chain length leads to better coverage performance of the standard confidence intervals. The effect this variation has on the coverage performance of the smoothed intervals (at fixed u) is inconclusive.
(4) Overall, it appears that the best coverage performance (always higher than the nominal value 90%) corresponds to the smallest value allowed for the u, $u = 0.5$.

REFERENCES

[1] Athreya, K.B., & Fuh, C.D. (1992). Bootstrapping Markov chains. In R. LePage & L. Billard (Eds.), *Exploring the Limits of Bootstrap* (pp. 49-64). New York: John Willey & Sons.
[2] Balescu, R. (1975). *Equilibrium and Nonequilibrium Statistical Mechanics*. New York: John Wiley & Sons.
[3] Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Chicago: The University of Chicago Press.
[4] Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
[5] Efron, B. (1979). Bootstrap method: another look at the jacknife. *Annals of Statistics 7*, 1-26.
[6] Efron, B. (1987). Better bootstrap confidence intervals (with comments). *Journal of the American Statistical Association, 82*, 171-200.
[7] Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.
[8] Fienberg, S.E., & Holland, P.W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association, 68*, 683-691.
[9] Guerra, R., Polansky, A.M., & Schucany, W.R. (1997). Smoothed Bootstrap confidence intervals with discrete data. *Computational Statistics and Data Analysis, 26*, 163-176.
[10] Hoyle, L. (1968). *The Influenza Viruses*. New York: Springer-Verlag.
[11] Kulperger, R.J., & Prakasa Rao, B.L. (1989). Bootstrapping a finite state Markov chain. *The Indian Journal of Statistics, 51 (A, Pt. 2)*, 178-191.

[12] Lippe E., De Smidt J.T., & Glenn-Levin D.C. (1985). Markov models and succession: a test from a heathland in the Netherlands. *Journal of Ecology, 73*, 775-791.
[13] Polansky, A.M. (1999). Upper bounds on the true coverage probability of bootstrap percentile method confidence intervals. *American Statistician, 53,* 362-369.
[14] Ross, S.M. (1993). *Introduction to Probability Models*. Boston: Academic Press.
[15] Serfling, R.L (1980). *Approximation Theorems in Mathematical Statistics*. New York: John Wiley and Sons.
[16] Titterington, D.M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics, 22*, 259-268.

Department of Statistics, University of New Mexico, Albuquerque, NM 87131